Introduction à la synthèse vocale

Simon Richard

Université de Montréal

Avril 2015

Qu'est-ce que la synthèse vocale ?

Tâche

Produire un signal acoustique à partir d'une séquence de mots

Sous-tâches

1. Convertir le texte en représentation phonétique (text analysis)
2. Convertir la représentation phonétique en signal (waveform synthesis)

Usages

- agent conversationnel
- application non-conversationnelle (p. ex. dispositif de lecture pour aveugles)
- communication améliorée et alternative

Désambiguïsation du point

He said the increase in credit limits helped B.C. Hydro achieve record net income of about $1 billion during the year ending March 31.

Cousins, however, was insistent that all debts will be collected : "We continue to pursue monies owing and we expect to be paid for electricity we have sold."

The group included Dr. J. M. Freeman and T. Boone Pickens Jr.

On peut utiliser une méthode d'apprentissage supervisé ; on indique où se trouvent les frontières de phrase dans un corpus d'entraînement pour entraîner le classificateur

On peut aussi se baser sur la probabilité qu'un mot soit en début ou fin de phrase pour déterminer si ce mot est EOS (end-of-sentence) ou pas

Normalisation du texte

Expansion des mots non standards

Les mots non standards (abréviations, acronymes, etc.) doivent être désambiguïsés et convertis en séquences de mots

Les nombres sont des mots non standards très ambigus :

Désambiguïsation du nombre 1750

seventeen fifty (comme dans *The European economy in 1750*)
one seven five zero (*The password is 1750*)
seventeen hundred and fifty (*1750 dollars*)
one thousand seven hundred and fifty (*1750 dollars*)

Pour gérer les mots non standards on doit : 1) faire la tokénisation pour identifier les mots non standards potentiels ; 2) les classifier ; 3) les convertir en séquences de mots selon la classification

Des expressions régulières suffisent pour certaines classes ; pour les autres, on fait recours à de l'apprentissage machine

Normalisation du texte

Désambiguïsation des homographes

Les homographes sont des mots qui partagent une même graphie mais sont prononcés différemment (p. ex. *couvent* (V) et *couvent* (N))

La plupart des homographes peuvent être désambiguïsés à l'aide d'un POS tagger car ils n'appartiennent pas à la même partie du discours ; ceux qui ne peuvent être désambiguïsés de cette manière sont souvent ignorés

P. ex. est-ce que *bass* (N) réfère au poisson ou à l'instrument ?

Analyse phonétique

Pourquoi est-ce difficile ?

En français québécois, la lettre d peut représenter plusieurs sons :

- d dans donner
- dz dans direction

En russe, certaines voyelles non-accentuées sont prononcées différemment :

- спас**и**бо (spas**i**bo) se prononce spasiba
- квеб**е**к (kveb**e**k) se prononce kvibek

L'accentuation n'est pas prévisible à partir de la forme du mot !

Les homographes couvent (V) et couvent (N) s'écrivent de la même façon mais se prononcent différemment

Analyse phonétique

Dictionnaires de prononciation

Il existe plusieurs dictionnaires de prononciation de mots-formes anglais, par exemple PRONLEX et CELEX2 qui sont distribués par le Linguistic Data Consortium (LDC) et CMUdict qui est open source

Il y a aussi UNISYN qui a été conçu spécialement pour la synthèse et qui a quelques fonctionnalités intéressantes (voir diapo ? ?)

Mais :

- de nombreux mots sont manquants (p. ex. les noms propres)
- l'anglais a une morphologie relativement pauvre comparé à d'autres langues comme le français, le russe, etc.
- les mots ne sont pas isolés lorsqu'on parle ; comment fait-on pour rendre compte de la liaison en français ?

Analyse phonétique

Conversion graphème → phonème (g2p)

On peut utiliser des règles pour attribuer une prononciation aux mots qui ne figurent pas dans notre dictionnaire :

$$\mathrm{d} \rightarrow \mathrm{dz} \; / \; \_ \left\{ \begin{matrix} \mathrm{u} \\ \mathrm{i} \end{matrix} \right\}$$

Mais on favorisera plutôt l'apprentissage machine :

$$\hat{P} = \underset{P}{\operatorname{argmax}}\, p(P|L)$$

Où L est une séquence de lettres, P une séquence de phones et $\hat{\mathrm{P}}$ la séquence la plus probable

De l'information supplémentaire peut aider ; par exemple, connaître le mot suivant permet de rendre compte de la liaison en français

Analyse phonétique

Simuler un accent

La phonologie est une branche de la linguistique qui traite de l'organisation des sons et de leur interaction

Elle nous permet de décrire des régularités de la langue, comme l'alternance entre `d` et `dz` en français québécois :

$$\texttt{d} \rightarrow \texttt{dz} \;/\; \_ \begin{Bmatrix} \texttt{y} \\ \texttt{i} \end{Bmatrix} \qquad \texttt{d} \rightarrow \texttt{dz} \;/\; \_ \begin{bmatrix} \text{-cons} \\ \text{+haut} \\ \text{-arr} \end{bmatrix}$$

Cette règle indique que `d` devient `dz` lorsque suivi du son `y` ou `i` ; `t` devient `ts` dans le même contexte

N.B. Le symbole `y` représente ici la voyelle qu'on retrouve à la fin du mot `barbu` et appartient à l'alphabet phonétique international (API)

Les traits distinctifs

Le système des traits distinctifs a été introduit par Chomsky & Halle en 1968 dans *The Sound Pattern of English*

<table>
<tr><td rowspan="5">-cons</td><td rowspan="5">+son</td><td rowspan="2">+haut</td><td>j</td><td>ɥ</td><td></td><td>w</td><td rowspan="2">-bas</td></tr>
<tr><td>i ɪ</td><td>y ʏ</td><td>ɨ ʉ</td><td>u ʊ</td></tr>
<tr><td rowspan="2">-haut</td><td>e ɛ</td><td>ø œ</td><td>ʌ ə</td><td>o ɤ</td></tr>
<tr><td>ɜ æ</td><td></td><td>ɐ a ɑ</td><td>ɒ</td><td>+bas</td></tr>
<tr><td>-lab</td><td>+lab</td><td>-lab</td><td>+lab</td></tr>
<tr><td colspan="2">-arr</td><td colspan="2">+arr</td></tr>
</table>

Les traits permettent de définir des classes de sons et de référer à ces classes à l'intérieur de règles phonologiques

Analyse prosodique

On se retrouve avec une séquence de phones ; faute d'ajouter de l'information prosodique, la synthèse vocale sera peu naturelle

La prosodie comprend l'information suprasegmentale, soit l'information non phonologique comme l'intonation, la hauteur, etc.

L'accent est important car il permet de distinguer des mots autrement identiques dans des langues comme l'italien :

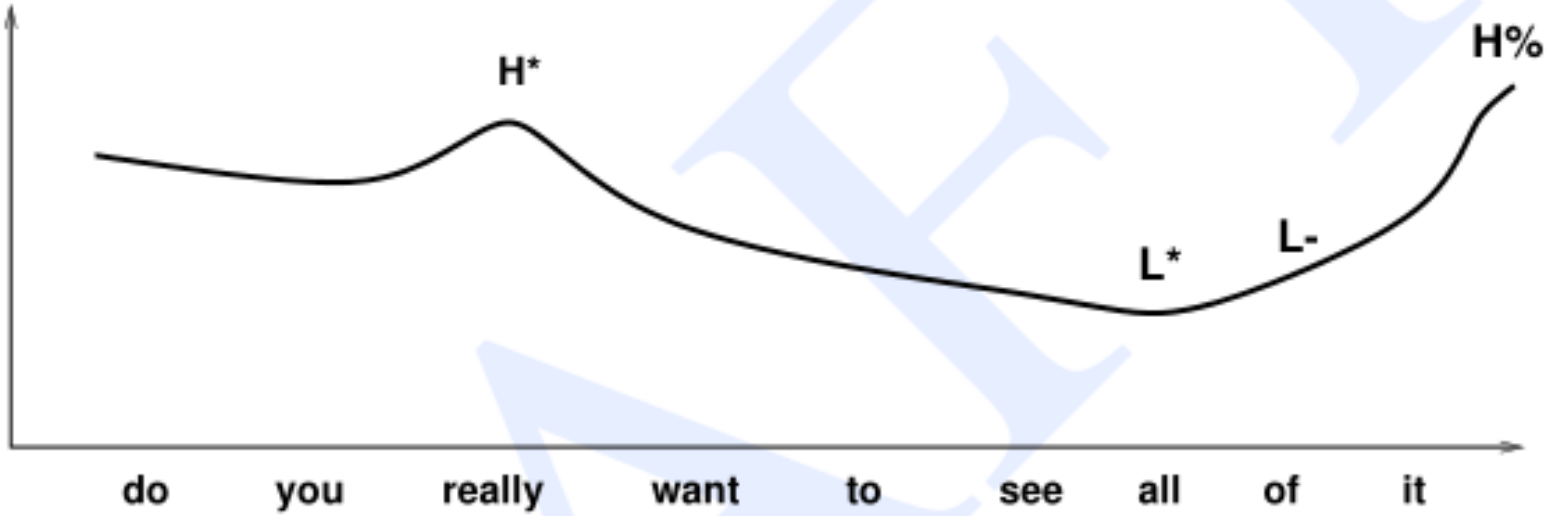
ancora veut dire ancre, alors ancora veut dire encore

L'intonation permet de distinguer une déclarative d'une interrogative
On attribue donc une intonation et une durée à chaque phone ; nous ne nous attarderons pas aux méthodes employées à cette étape

Analyse prosodique

Le résultat

				H*												L*		L- H%			
do		you		really				want				to		see		all		of		it	
d	uw	y	uw	r	ih	l	iy	w	aa	n	t	t	ax	s	iy	ao	l	ah	v	ih	t
110	110	50	50	75	64	57	82	57	50	72	41	43	47	54	130	76	90	44	62	46	220



Générer le signal acoustique

Le résultat de l'analyse textuelle est une séquence de phones auxquels sont attribuées une durée et une intonation

On ne peut pas se contenter de concaténer (mettre bout à bout) les phones ; le signal acoustique qui correspond à t n'est pas le même devant a et s et le résultat serait peu naturel

Deux solutions :

- simuler la coarticulation (c-à-d la transition entre les phones)
- utiliser des unités plus larges (p. ex. des diphones, syllabes, mots)

Générer le signal acoustique

Les diphones

On préfère généralement les diphones ; ceux-ci débutent au milieu du premier phone et terminent au milieu du second phone ; les phones sont moins stables à leurs « extrémités »

On en compte $\sim$1000 à 2000 par langue, ce qui est plus économe que de stocker des triphones, etc ; pour n phones on compte n^2 diphones possibles (1849 en anglais), mais les combinaisons ne sont pas toutes possibles

Générer le signal acoustique

La synthèse concaténative à l'aide de diphones (diphone waveform synthesis)

Entraînement :

- Enregistrer quelqu'un qui prononce chaque diphone
- Stocker les diphones dans une base de données

Synthèse :

- Récupérer les diphones correspondant à la séquence voulue
- Concaténer les diphones en modifiant légèrement leurs frontières
- Modifier le signal pour obtenir la prosodie voulue

Désavantages :

- La manipulation du signal laisse des artéfacts
- La coarticulation ne dépend pas toujours du phone voisin seulement

Générer le signal acoustique

La synthèse concaténative à sélection d'unité (unit selection synthesis)

On préfère cette approche à la précédente

Différences :

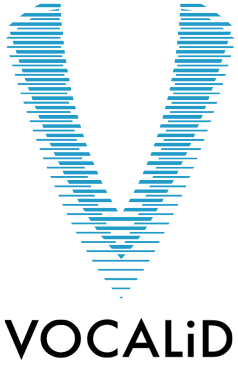
- On stocke plusieurs copies de chaque diphone alors qu'on en stocke une seule en diphone waveform synthesis
- On manipule peu ou pas les éléments concaténés alors qu'on utilise différents algorithmes pour modifier la prosodie des diphones en diphone waveform synthesis

Pour chaque diphone, on sélectionne la copie dont la prosodie demande le moins de manipulation et dont les frontières sont compatibles avec les diphones voisins

On attribue à chaque copie un coût et on choisit la moins coûteuse

Un exemple d'application

VocaliD - http://www.vocalid.co



Nous avons parlé brièvement des applications AAC au début de cette présentation

Dans bien des cas, la même voix est utilisée pour un homme adulte (p. ex. Stephen Hawking) et pour une fillette de 7 ans

VocaliD recueille des dons de voix et crée des voix personnalisées aussi proches de l'identité vocale du bénéficiaire que possible avec une technique de « voice blending » à la fine pointe de la technologie

Références

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York, NY : Harper & Row.

Jurafsky, D., & Martin, J. H. (1999). *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition* (1e ed.). Upper Saddle River, NJ : Prentice Hall.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition* (2e ed.). Upper Saddle River, NJ : Prentice Hall.